Right to an Explanation Considered Harmful

Right to an Explanation Considered Harmful

Andy Crabtree [1] · **Lachlan Urquhart** [2] · **Jiahong Chen** [3]

Abstract

*Lay and professional reasoning has it that newly introduced data protection regulation in Europe – GDPR – mandates a 'right to an explanation'. This has been read as requiring that the machine learning (ML) community build 'explainable machines' to enable legal compliance. In reviewing relevant accountability requirements of GDPR and measures developed within the ML community to enable human interpretation of ML models, we argue that this reading should be considered harmful as it creates unrealistic expectations for the ML community and society at large. GDPR does not require that machines provide explanations, but that data controllers – i.e., **human beings** – do. We consider the implications of this requirement for the 'explainable machines' agenda.*

1. Introduction

" … the paucity of critical writing in the machine learning community is problematic … as machine learning continues to exert influence upon society, we must be sure that we are solving the right problems … Thus, we believe that such critical writing ought to have a voice at machine learning conferences." (Lipton, 2016)

The title of this paper reflects a longstanding tradition in computer science initiated by Edsger Dijkstra in 1968 who wrote 'Go To Statement Considered Harmful', a critique of existing programming practices that eventually led the programming community to adopt structured programming (Dijkstra, 1968). Since then, titles that include the phrase 'considered harmful' signal a critical essay that advocates change. In this case the change in question concerns treating GDPR as mandating a 'right to an explanation', that might be met through the ongoing development of machine learning (ML) methods of interpretability, as *problematic* and indeed *harmful* to the ML community and society at large.

A cursory search of the Internet using the terms 'GDPR right to explanation' reveals that lay and professional reasoning widely read that new data protection regulation in Europe mandates a right to an explanation when their data is processed by algorithmic machines. Given GDPR's territorial scope (see Article 3 GDPR, 2018), this right is effectively read as being globally applicable insofar as automated decision-making touches EU citizens *or* their data wherever it or they may reside. This reading has been seen by many commentators, practitioners and scholars to mean that we need to develop 'explainable machines' that can account for algorithmic decision-making. Taglines such as the following are not uncommon and feed the hype about the implications of GDPR for AI and machine learning.

> **AI Will Have to Explain Itself**
> The need for Explainable AI is being driven by upcoming regulations, like the European Union's General Data Protection Regulation (GDPR), which requires explanations for decisions ... Under the GDPR, there are hefty penalties for inaccurate explanations – making it imperative that companies correctly explain the decisioning process of its AI and ML systems, every time. (Zoldi, 2018)

The perceived requirements of GDPR appear to segue neatly on the face of it with the ML community, which has been trying to find ways to render complex machine learning models 'interpretable' for over thirty years. Naturally we think this a laudable aim, for regardless of legal requirements people still have need to understand the machines they build and use. However, we are not convinced that ML methods of interpretability will meet the requirements of GDPR. Indeed, we argue that the perceived alignment between interpretability in ML and legal explanation should be considered harmful, insofar as it creates unrealistic expectations of what the ML community can deliver and what society at large can expect from algorithmic machines.

[1] School of Computer Science, University of Nottingham, UK.

[2] Edinburgh Law School, University of Edinburgh, UK.

[3] Jiahong Chen, Horizon Digital Economy Research Institute, University of Nottingham, UK.

Contact: andy.crabtree@nottingham.ac.uk

In what follows we explicate the grounds of our assertion in considering relevant accountability requirements of GDPR, particularly that it is the 'data controller', i.e., the party who determines the means and purposes of data processing, who is legally obligated to provide an explanation, **not** a machine and certainly **not** an algorithm. We also consider the kinds of explanation enabled by ML methods of interpretability and their resonance or 'fit' with salient accountability requirements of GDPR. We conclude in considering what is required of an explanation by GDPR and the societal imperative occasioned by the widespread introduction of inscrutable, non-intuitive algorithmic machines into everyday life.

2. The Right to an Explanation?

The accountability requirements of GDPR oblige data controllers to put in place effective policies and mechanisms to *demonstrate* that the processing of personal data is in compliance with GDPR (Urquhart et al., 2018). In this respect GDPR (2018) mandates in Article 13 (information to be provided when data are collected from the data subject), Article 14 (information to be provided when data have not been obtained from the data subject), and Article 15 (right of access by the data subject) that certain information must be provided to the data subject. This includes *"the existence of automated decision-making, including profiling, referred to in Article 22 … and meaningful information about the logic involved … the significance and the envisaged consequences of such processing for the data subject (ibid.)."* Article 22 (automated individual decision-making, including profiling) mandates the conditions under which automated-decision making may take place.

The *"meaningful information"* clause has led many commentators, practitioners and scholars to conclude that GDPR therefore mandates a right to an explanation with respect to the decisions made by algorithmic machines. While some contest the idea (see Wachter et al. 2017), our concern is that the nature of the explanation required by GDPR is commonly misunderstood and that this misunderstanding is harmful to the machine learning community and society at large insofar as it creates unrealistic expectations for both parties. Legal-tech scholars Edwards and Veale (2017) hint at the nature of the misunderstanding in accrediting it to a short paper written by Goodman and Flaxman (2016), who interpret the meaningful information clause as a *"requirement [that] prompts the question: what does it mean, and what is required, to explain an algorithm's decision?"*

Now when we look at GDPR for what it may say about what it means to explain an algorithm's decision we find as others have found (e.g., Wachter et al., 2017; Selbst and Powles, 2017; Edwards and Veale, 2017; Floridi et al., 2017) that there is no mention of explanation in any Article – and only Articles create legally binding obligations – but let us set that aside for the moment. What we do find, as stated above, is a requirement that the 'data subject' – i.e., the person whose data is to be, is being or has been processed – be informed of the logic involved in automated processing, as well as the significance and the envisaged consequences of such processing. So at the outset the meaningful information required is not information, per Goodman and Flaxman (2016), that explains an algorithm's decision but rather, as Selbst and Barocas (2018) put it, *"a functional description of the model [or] the rules governing decision-making"*.

Furthermore, this functional description of the model or rules governing decision-making, which should be written in *"clear and plain language"* (Recital 39 GDPR, 2018), is prospective in nature and applies to *all* potential cases of automated decision-making, not just the particular case to hand. *"The most important aspect of this type of explanation is that it is concerned with the operation of the model in general, rather than as it pertains to a particular outcome* (Selbst and Barocas, 2018)." So whatever an explanation might amount to as mandated by Articles 13 and 14 it has nothing to do with explaining an algorithm's decision.

Rather, in the case of Article 13 explanation is about providing *ex ante* information that allows the data subject to make an informed choice as to whether or not they wish to engage in and subject their data to the kind of automated decision-making offered by the data controller when entering into a contract or consenting to such processing. In the case of Article 14, explanation is about providing sufficient information to allow the data subject to *"vindicate her other substantive rights under the GDPR and human rights law"* (ibid.), including the right to obtain from the controller confirmation as to whether or not personal data concerning him or her are being processed and, if they are, to exercise their right of access and request from the controller rectification or erasure of personal data or restriction of processing of personal data or to object to such processing, as per Article 15.

As Edwards and Veale (2017) note in referring to the right of access, Article 15 implies that data has been collected and processing has begun or taken place. It therefore appears to promise a right to an explanation *ex post* and that "*tailored knowledge about specific decisions"* (ibid.) should be provided to the data subject. Article 15 thus appears to speak to the kind of explanation invoked by Goodman and Flaxman. However, as Edwards and Veale observe, Article 15 *"has a carve out in the recitals, for the protection of trade secrets and IP."* Thus Recital 63 (GDPR, 2018) states that the right of access *"should not adversely affect the rights or freedoms of others, including trade secrets or intellectual property and in particular the copyright protecting the software."* So it would appear that the right to an explanation as construed by Goodman and Flaxman exists, but is limited.

However, that would be to concede too much. As noted above, there is no mention of a right to an explanation in any Article within GDPR. Where we do find mention of it is in the Recitals, which support interpretation of an Article's meaning and requirements. Recital 71 is tied to Article 22

(see ICO, 2018) and enables interpretation of Article 22(3) (GDPR, 2018) in particular: *"In the cases referred to in points (a) [automated processing based on performance of a contract] and (c) [or consent] … the data controller shall implement suitable measures to safeguard the data subject's rights and freedoms and legitimate interests, at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision."*

What the *"right to obtain human intervention"* means or involves is further clarified by Recital 71, which states that *"any form of automated processing of personal data … should be subject to suitable safeguards, which should include specific information to the data subject and the right to obtain human intervention, to express his or her point of view, **to obtain an explanation of the decision reached after such assessment** and to challenge the decision."* So GDPR clearly mandates a right to an explanation with respect to automated decision-making, but is it as Goodman and Flaxman (2016) propose an explanation of an algorithm's decision? In explicating the role of explanation with respect to AI and the law, Budish et al. (2017) note that *"when we talk about an explanation for a decision … we generally mean the **reasons or justifications** for that particular outcome, rather than a description of the decision-making process."* The kind of explanation required by GDPR is not an account of how and algorithm arrived at a decision then, but an account that justifies that decision and which enables the data subject and others, including regulatory authorities and lawyers, to evaluate the reasonableness of the decision. (Selbst and Barocas, 2018).

So it is not that the right to an explanation of an algorithm's decision is merely limited by IP and copyright under Article 15. It is that where the right to an explanation is explicitly mandated then it is **not** concerned to explain how an algorithm arrived at a decision, but rather to justify, make reasonable and legally defensible that decision. Misreading GDPR and misunderstanding the kind of explanation required creates what Edwards and Veale (2017) call a "transparency fallacy" that focuses attention, mistakenly, on the inner workings of the algorithmic machine. The right to an explanation is not concerned with those workings but ensuring the rights and freedoms of the data subject are respected and protected. In this respect Article 22 introduces further difficulties insofar as its requirements *only* apply when the processing of personal data has been performed *solely* by automated means and produces *legal* effects (e.g., it impacts a person's legal status or their legal rights) or has consequences that *significantly* affect the data subject's circumstances, behaviour or choices (e.g., refusal of credit).

Where does that leave us then? GDPR clearly mandates a right to an explanation with regards to automated decision-making. However, that right is not do with explaining an algorithm's decision. Rather what is mandated requires a) that a general *ex ante* explanation of the logic, significance and envisaged consequences of the automated decision-making model be provided to the data subject and b), **if** the decision-making is wholly automated, and **if** it produces legal or significant effects, that an *ex post* explanation justifying the specific decision or decisions arrived at by an algorithmic machine be provided to the data subject, **if** the data subject requests it and **if** that request does not affect the data controller's rights and freedoms or the rights and freedoms of others who enable the controller's processing operation (e.g., software providers). However, the IP/copyright carve out cannot be used as grounds *not* to provide *any* explanation to the data subject (see Recital 63 GDPR, 2018).

In saying this we are aware that Selbst and Powles (2017) argue the *ex ante* / *ex post* distinction *"falls apart"* on analysis. However, it is important to appreciate that what falls apart is the particular distinction put forth by Wachter et al. in arguing against the right to an explanation as construed by Goodman and Flaxman (2016). For Wachter et al. (2017), GDPR *only* mandates *ex ante* explanation; the right to *ex post* explanation concerning specific decisions is *"incorrectly attributed to Article 22(3) – which only features in Recital 71."* The argument put forth by Wachter et al. is that the interpretation of GDPR offered by Goodman and Flaxman is therefore incorrect. Selbst and Powles (2017) agree but on different grounds and argue that in trivialising Recital 71, Wachter et al. *"ignore [its] positive value";* that it is *"not meaningless, and has a clear role in assisting interpretation and co-determining positive law."* A point underscored by the UK supervisory authority (ICO, 2018).

Nonetheless, there would appear to be no **general** provision mandating the explanation, *ex post*, of specific decisions arrived at by algorithmic machines in GDPR. *Ex post* explanations – and specifically justifications of specific decisions arrived at by algorithmic machines – are only required in certain circumstances as outlined above. As Edwards and Veale (2017) describe it, *"these certainly seem shaky foundations on which to build a harmonised cross EU right to algorithmic explanation."* Indeed, the foundations seem globally shaky given the territorial scope of GDPR. Shakier still is the idea that explanation should be furnished by a machine: *"the protection of natural persons should be technologically neutral and should not depend on the techniques used* (Recital 15 GDPR, 2018)." On the contrary, what is mandated, as stated in Article 22(3) and writ large in Recital 71 is the right for the data subject to obtain an explanation through ***"human intervention"***. It is people, not machines, that GDPR requires an explanation from. What then for explainable machines?

3. Enabling Human Intervention?

Now it might be argued that the right to obtain an explanation through human intervention does not negate efforts to build explainable machines. As Goodman and Flaxman (2016) put it, *"it is reasonable to suppose that any adequate explanation would, at a minimum, provide an account of how input features relate to predictions."* Indeed, we might think with

good reason that some account of how an algorithmic machine arrived at a decision would need to be incorporated into the occasioned provision of a justification. So there would seem to be great promise for those who are obliged to provided explanations on the particular occasions when they are required by GDPR, *and* for those who must assess the reasonableness of the explanations offered, in the methods of interpretability offered by the ML community.

In a survey of methods for explaining 'black boxes', Guidotti et al. (2018) describe explanation is an ***"interface"*** between humans and an automated decision-maker. We take it that this interface is not primarily concerned with algorithmic transparency or understanding, as Molnar (2018) puts, *"how the algorithm learns a model from the data and what kind of relationships it is capable of picking up. If you are using convolutional neural networks for classifying images [for example], you can explain that the algorithm learns edge detectors and filters on the lowest layers. This is an understanding of how the algorithm works,* ***but not of the specific model that is learned in the end and not about how single predictions are made****."* So while algorithmic transparency is important, it is not the principle problem. As Hughes (in Medsker, 2017) puts it, *"while there are some algorithms involved, machine learning often contains a great deal of inference."* Algorithmic transparency is necessary but not sufficient, as the *"inference engine"* or *"model"* (ibid.) produced by the algorithm requires further explanation.

Guidotti et al. (2018) identify two fundamental kinds of model explanation in the ML literature. One that focuses on **global interpretability** and explaining *"the whole logic of a model and … the entire reasoning leading to all the different possible outcomes"*, and another that focuses on **local interpretability** and explaining *"only the reasons for a specific decision."* Thus Guidotti et al. find two different orders of explanation at work in the ML literature – one that focuses on describing how 'black boxes' work, and the other on explaining the decisions they make – and these lead to two fundamental approaches to explanation: "*design of explanations"* and *"reverse engineering"* (ibid.). The former involves selecting and training a machine learning model that is considered to be intrinsically interpretable. The latter, and more common approach, focuses on furnishing a post hoc explanation of the decision arrived by an algorithmic machine.

Guidotti et al. go on to describe in detail a *"family of explanation problems"* and *"explanator methods"* found in the ML literature. The authors note, however, that despite the arsenal of methods available *"in the literature, very little space is dedicated to a crucial aspect: the model complexity. The evaluation of the model complexity is generally tied to the model comprehensibility, and this is a very hard task to address* (ibid.)." As Lipton (2016) reports, even the assumption that intrinsic approaches enable model comprehensibility, and thus interpretation and explanation, is problematic: *"neither linear models, rule-based systems, nor decision trees are intrinsically interpretable. Sufficiently high-dimensional models, unwieldy rule lists, and deep decision trees could all be considered less transparent than comparatively compact neural networks."* It would appear that global interpretability is very difficult to achieve.

Reverse engineering methods are not without their problems or dangers either. While post hoc methods may work well in explaining the distinction, for example, between wolves and huskies (Riberio et al., 2016), they may not be so effective in cases where models take a much larger set of features into account, in which case the complexity problem still applies. Interactive methods (e.g., Diakopoulos, 2016) are also potentially problematic insofar as they might placate rather than elucidate. As Selbst and Barocas (2018) put it, *"people could try to make sense of variations in the observed outputs by favouring the simplest possible explanation that accounts for the limited set of examples that they generated by playing with the system."* So while holding initial promise, significant challenges confront post hoc methods in dealing with and conveying model complexity and comprehensibility to human beings, just as they confront intrinsic methods.

However, more problematic for the purposes of this paper, the kinds of explanation offered by ML methods of interpretability do not segue as neatly as they might at first appear with the requirements of GDPR or the obligations of those who must meet its requirements or require information to satisfy them. While a globally interpretable account of a machine learning model may support explanation of the logic involved in automated decision-making, it does **not** account for the significance or consequences of it, just as a locally interpretable account does **not** justify the specific decision arrived at by an algorithmic machine. There are certain explanations that necessarily stand **outside** algorithmic decision-making and cannot be provided by methods of interpretability.

At the root of the matter here is the recognition, as Guidotti et al. (2018) note, that *"each community … provides a different meaning to explanation."* While the meaning of explanation is not settled in the machine learning community (see Lipton, 2016; Guidotti et al., 2018), it would appear to revolve around the notion of causality and explaining, as Goodman and Flaxman (2016) stress, *"****how*** *input features relate to predictions."* Molnar (2018) goes further, *"you do not want a human-style explanation, but rather a* ***complete causal attribution*** *[at least] you probably want a causal attribution when you are legally required to state all influencing features"* or as Zoldi (2018) puts it explanation needs to account for *"the decisioning* ***process****"*, not just the relationship between input and output features. Now we appreciate that the nature and level of causal account required to explain automated decision-making is debateable and far from settled, and will turn upon what needs to be explained on any occasion, but our point here is that when ML focuses on explanation then it seeks to provide a causal account of some kind that articulates how a decision was arrived at.

It might also be argued that the meaning of explanation is not settled in law either – see the competing accounts of Wachter et al. (2017) and Selbst and Powles (2017) on the right to an explanation in GDPR for example – but it too arguably revolves around a core notion, not of cause but of **reason**. Thus, at the root of the misunderstanding of what GDPR requires are **two very different orders of explanation**. Queloz (2017) hints at the difference in the following example: "*The rule 'if the signal is red, then stop', together with the fact that its antecedent is fulfilled, justifies the action, while giving the cause of [the] action would not justify it. Any actually performed transition from one proposition to another, or from thought to action, has a causal basis, a physiological realisation. Yet what justifies the transition is not that causal basis, but the normative and factual considerations that make the transition correct or incorrect.*"

What Queloz's example makes perspicuous is that a causal account **cannot** explain a person's decision to comply, or not, with the rule 'if the signal is red, then stop'. Describing the influx of photons into the eye, the physiological recognition of the colour red or the physiological transition from thought to movement of muscles and the skilful physical interaction between body parts and machine parts (brakes, clutches, gears, etc.) does not explain compliance with the rule. For that we must appeal to *"the normative and factual considerations that make the transition correct or incorrect"* – i.e., that the light is red, that we routinely stop at red lights, it is what we do in our culture, it is part and parcel of the business of driving in an orderly fashion, if we don't stop we may well be penalised, though emergency vehicles such as ambulances or fire engines or the police may go through them if they are on call, etc. In saying what any competent wide-awake adult knows, it is evidently the case that we are now in the business of giving reasons that account for and justify one's decision to stop and the decisions of others not to stop at the signal if it is red. **Reasons justify decisions** in the social world (Winch, 1958), not causes.

GDPR seeks to make automated decision-making an explainable feature *of the social world* and thus accountable to data controllers, data subjects, regulators, lawyers, judges, law-makers, etc. This is not to say that causes have no place in justifying decisions; it is to say that causes cannot justify decisions in themselves (Raz, 2011). As Queloz (2017) puts it, *"causes can be referred to in justifications, since a causal relation can hold between events even though the events are referred to under descriptions linking them in a justificatory relationship."* So, for example, cause of death by heart attack might be invoked to explain the failure of an aged gentleman to comply and to justify the no fault insurance claims of others affected by his tragic circumvention of the rule 'if the signal is red, then stop.' Causal accounts may be implicated in legal explanations, and indeed human explanations more generally, but they **do not suffice** in themselves. More is required and that of course is what GDPR demands, which rather limits the scope of machine learning methods of interpretability in enabling the kind of explanation required.

The idea that machine learning methods align or can align with the right to an explanation mandated by GDPR is essentially rooted in confusions in ordinary language and the double use of the word 'why'. On the one hand we routinely use it to ask questions about the causes of things and to elaborate how they came about. On the other, we routinely use it to inquire into persons reasons for doing things, including the ways they were done, and to thereby justify them. As Wittgenstein (1992) points out, ordinary language often *"bewitches"* us and gets us into conceptual trouble, in this case giving rise to the misunderstanding that a causal account is sufficient to deliver the kind of explanation required by GDPR, namely an *ex ante* explanation of the automated decision-making model and its significance and consequences for the data subject, and an *ex post* explanation justifying the specific decision or decisions arrived at by an algorithmic machine.

4. Human-Machine Accountability

Even if an adequate causal account can be arrived at, ML methods of interpretability can but play a **limited role** in the provision of a legally defensible explanation. As Selbst and Barocas (2018) point out, methods of interpretability *"cannot address why decisions happen to be made that way or whether the decisions are justifiable."* The authors go on to elaborate a distinctive order of reasoning to which causal accounts are generally held accountable by the legal profession and human beings more generally, using the following example from the ML literature by way of explication. *"Caruana et al. (2015) … discovered that a model trained to predict complications from pneumonia had learned to associate asthma with a reduced risk of death. To anyone with a passing knowledge of asthma and pneumonia, this result was obviously wrong."*

It turns out that the model was not wrong, but it did not make perspicuous that asthma patients pay much closer attention to their breathing than non-asthma patients and thus pre-empt the onset of fatal respiratory disease. The problem with the model – like many other ML models – is that it is not available to intuition which, as Selbst and Barocas (2018) describe it, is *"the bridge by which we go from explanation to normative assessment … in evaluating machine learning models … through a broad range of experiences, typically described as 'common sense'."* The justification of automated decision-making in the social world thus turns on **the availability of a causal account to common sense reasoning** and with it normative assessment of the reasonableness of an algorithmic model's decision.

Now one might balk at the suggestion that legally defensible explanations are dependent upon common sense reasoning. However, as sociologist Harvey Sacks found during his time as a law student at Yale there is more to the law than statute. Confronted one day by a problem in case law as to whether or not a person on the ground was entitled to recover damages incurred from the overflight of his property by an airplane, it

was suggested that no damages could be collected if a plane was being piloted in a proper manner. The ensuing legal argument that occupied Sacks and his cohort turned on the definition of what was proper? *"What if it were flying at 2,000 feet? At 1,000 feet? At 250 feet? At 5 feet? Sacks reported that when the last of these proposal was offered, it was dismissed as 'unreasonable', as frivolous, as violating the canons of 'common sense' … he pointed out that could have as well have been said about the penultimate one, but wasn't. What struck him then … was that 'legal reasoning' … was constrained by an infrastructure of so-called 'common sense' which was entirely tacit and beyond the reach of argument, while* ***controlling it****"* (Schegloff, 1992).

The good news for the ML community, as Selbst and Barocas (2018) also point out, is that common sense reasoning is often *"flawed."* As the above example regarding asthma and pneumonia makes perspicuous, intuition is not well equipped to deal with decisions that run counter to common sense, and this goes to the nub of the practical (if not legal) need we have for 'explainable AI'. As Gunning (2018) puts in outlining DARPA's XAI programme, *"explainable AI – especially explainable machine learning – will be essential if future warfighters are to understand, appropriately trust, and effectively manage an emerging generation of artificially intelligent machine partners."* Of course it is not only 'warfighters' who need to understand, trust, and manage artificially intelligent machine partners, ultimately we are all going to have to be able to do that, but the limitations of common sense reasoning surface the fundamental challenge confronting efforts to explain algorithmic machines: that the value of machine learning lies in it finding patterns that go **well beyond** human intuition.

There is a **double-bind** at work here which underpins our argument as to why the right to an explanation should be considered harmful. On the one hand legally defensible explanations require that causal accounts of automated decision-making be made accountable to common sense reasoning to enable normative assessment of the reasonableness of, and thereby justify decisions arrived at, by algorithmic machines; and on the other the very virtue of ML increasingly lies in its complexity and consequent unavailability to common sense account. It is not only that algorithmic machines may be 'inscrutable', meaning as the UK's Chief Scientific Advisor, Mark Walport (2018) puts it, that *"we can't know the precise structure and workings of algorithms that evolve continuously by a process of machine learning"*, it is that the decisions they arrive at are also in a great many cases **non-intuitive**.

We are reminded of the philosopher Ludwig Wittgenstein's elliptic remark, *"If a lion could speak, we could not understand him"*, which essentially means that lions are incapable of making their reasoning available to us as they have an entirely different *"form of life"* (Wittgenstein, 1992). Yet we co-exist and, just as with lions, then so we are going to have to learn to accommodate inscrutable, non-intuitive machines in our world. It is for this reason that we say that the right to an explanation should be considered harmful, as it blinds us to the need to consider what the **accommodation** of algorithmic machines might turn upon, how we might understand it, and what might be involved in enabling it. Instead, preoccupation with the right to an explanation creates unrealistic and unrealisable expectations that the ML community can enable it through methods of interpretability, and that society at large can duly expect everyday life to be populated by explainable machines. Both parties will be disappointed, and the latter may well become seriously disillusioned to adverse effect on the uptake of AI.

To be clear, the right to an explanation mandated by GDPR does not require an explanation of how an algorithmic machine arrived at a decision. It requires a) that an *ex ante* and generic account describing the logic, significance and consequences of an automated decision-making model be provided to the data subject, and b) in certain circumstances when the decision-making is solely automated, that an *ex post* account providing a justification of specific decisions arrived at be provided by a human being (the data controller) if the data subject requests it. ML methods of interpretability cannot deliver on this as those methods trade on causal accounts, on explaining *how* an automated decision was made, but not why it happened to be made in the way that was or whether the decision itself is justifiable. These matters necessarily sit outside of the machine. Hence, ML methods of interpretability can only play a limited role in explaining automated decision-making.

That said, algorithmic machines are increasing complex and commensurately inscrutable. As Knight (2017) puts it, "*You can't just look inside a deep neural network [for example] to see how it works. A network's reasoning is embedded in the behaviour of thousands of simulated neurons, arranged into dozens or even hundreds of intricately interconnected layers. The neurons in the first layer each receive an input … and then perform a calculation before outputting a new signal. These outputs are fed, in a complex web, to the neurons in the next layer, and so on, until an overall output is produced. Plus, there is a process known as back-propagation that tweaks the calculations of individual neurons in a way that lets the network learn to produce a desired output."* It may well be, then, that obtaining a causal account is deeply problematic, though we note that such an account is not necessarily required by GDPR; what is required is *justification* of a decision arrived at by automated means.

However, that too may be problematic insofar as automated decision-making is non-intuitive and rests on relationships that defy intuition and resist comprehension. So even if it is possible to obtain a causal account and to make the inner workings of the machine transparent, we may **not** understand it and not **be able to** understand it. The right to an explanation is beguiling. It seems to segue well with ML's interests in interpretability, but it doesn't. It sets up unrealisable expectations for the ML community – it will never be able to

deliver justifiable explanations, in contrast to causal explanations – and it sets up unrealisable expectations for society at large insofar as causal explanations may be very difficult, if not impossible, to obtain and even if they can then they may simply not be intelligible to human beings. As Mittelstadt (2016) puts it, algorithmic machines are often *"epistemically inaccessible."* Add to that, that it is reasons for, not causes of, decisions that people ordinarily seek when seeking explanations and the limited role of ML methods of interpretability in explaining automated decision-making seems decidedly brittle.

Preoccupation with the right to an explanation diverts attention from the fundamental need to accommodate inscrutable, non-intuitive machines in everyday life. If we are to address and understand the challenges of accommodation, we need to move outside the 'black box' and beyond internal measures of accountability that focus on accuracy, verification, and robustness (Chen 2018). And we need to move beyond ethical concerns that seek to eliminate bias, both in training and deploying algorithmic machines, and ensure privacy in data processing. This is not to say that ethical accountabilities implicated in automated decision-making are not important and should not be addressed, but that there is more to accommodating algorithmic machines in the social world than these *machine-oriented* concerns provide for or can provide for.

An additional layer of decisions is made by those who develop, deploy and use algorithmic machines, which occasions the need for **external** measures of accountability. As Selbst and Barocas (2018) put it, *"There is a set of explanations internal to the operation of the box itself, and a set of explanations about the design of the system and how the system will be used, that by necessity are external … When we seek to evaluate the justifications for decision-making that relies on a machine learning model, we are really asking about the institutional and subjective process behind its development … Evaluating models in a justificatory sense means comparing the choices made by the developers against society's broader normative priorities, as expressed in law and policy."*

Understanding what is involved in justifying automated decision-making leads to explanations outside the black box. Thus Wachter et al. (2018) propose *"counterfactual explanations"* that go beyond *"an attempt to convey the internal state or logic of an algorithm … [to] describe a dependency on the external facts that led to that decision"*, e.g., that credit wasn't given because the applicant doesn't earn enough. Selbst and Barocas (2018) propose *"algorithmic impact statements"* (AIS), i.e., documents that explain the choices about the decision-making model, how data was collected, the features that were and were not considered, and the anticipated effects of the automated decision-making process. And Casey et al. (2018) note that the right to an explanation mandates "*a general form of oversight"* that turns upon *"data auditing methodologies"*, particularly data protection impact assessments or DPIAs (Article 35 GDPR, 2018; A29WP, 2018; A29WP, 2017). Whatever way you cut it, more is required by the right to an explanation than ML can deliver.

5. The Limits of Explanation

Whether focused on internal or external accountabilities it would appear that there is consensus, at least, at the current moment in time that machine learning can be accommodated in everyday life through the ethical-legal-technical or ELT matrix.*"[The question of] whether and how algorithmic decision-making can be conducted in a 'transparent' or 'accountable' way, and the scope for decisions made by an algorithm to be fully understood and challenged … needs to be understood from an ethical, legal, and technical view."* (Floridi et al., 2017). We would suggest, however, that the **social imperative** of accommodating inscrutable and non-intuitive machines in everyday life cannot be adequately addressed in this way. The matrix assumes that citizens are Rational actors, who will readily accept AI because it is *asserted* ethical and made internally and externally accountable in various *possible* ways to the law.

We very much doubt that common sense reasoning will be quite so accommodating. One only needs to read texts such as Pasquale's *Black Box Society: The Secret Algorithms that Control Money and Information* (2015) to feel the palpable mistrust that permeates public discourse around algorithmic machines, and recent turns in Western politics rather underscore the fact that Rationalism is not a driving force in everyday life no matter how much one might want it to be. There is pressing need to move beyond the ELT matrix and consider ML **in context** and what it means to live with inscrutable, non-intuitive machines in everyday life. We need to get to grips with what happens **when explanations come to end** and all we can say, to borrow from the philosopher Norman Malcolm (1986), is *"this is what they do"*.

The ML literature makes it clear that what algorithmic machines do will not always be intelligible and explainable to us. There will also be problems, things will breakdown and go wrong; after all, no technology is infallible. 2018 brought with it the first death of a pedestrian, Elaine Herzberg, by an autonomous vehicle (Levin and Carrie, 2018). While investigation into this tragic incident sought to explain what happened and why things went wrong, and will seek do so in each and every case in the future, explanation is not the key challenge in accommodating AI. As Walport (2017) hints at, **social acceptability** is paramount, and this is not something we are going to determine through the ELT matrix alone. Rather, we need to understand the social expectations and concerns that ordinary people entertain about AI as seen from the perspective of their everyday lives, which would appear to have very little to do at first glance with what goes on inside the black box (Nilsson et al. 2018).

Broader **societal representation** is required to enable us to understand and address the challenges that are involved in *"domesticating"* AI (Lie and Sørensen, 1996) and making the technology at home in a socially organised world. The ELT matrix, preoccupied as it is with explainable AI, kicks this critical concern into the long grass. As the Royal Society (2017) puts it, *"As these new capabilities for computer systems become increasingly mundane, our relationship with – and expectations of – the technologies at hand will evolve. This raises questions about the long-term effects upon – or expectations from – people who have grown up with machine learning systems and smart algorithms in near-ubiquitous usage from an early age."*

As AI exerts increasing influence upon society we urgently need approaches that are capable of engaging with the mundane existential reality of AI here and now. While the preoccupation with explainability reaches beyond the ELT matrix (Abdul et al., 2018) alternative narratives are beginning to emerge in the field of Human Computer Interaction, where we find empirical studies of intelligent systems in use (e.g., Porcheron et al., 2018), and design approaches that seek to elicit the acceptability challenges confronting the widespread adoption of future and emerging technologies (e.g., Lindley et al., 2017) and to enable the appropriation of AI into everyday life (Kuijer and Giaccardi, 2018). Clearly a great deal more is required to get to grips with the **social imperative of accommodation and the commensurate need to broaden societal representation** than is provided by the ELT matrix, suffice to say here that the turn to the social in ML will turn not only on recognising the misdirection created by lay and professional readings of GDPR, but with it the limits of explanation itself.

6. Conclusion

That GDPR mandates a right to an explanation concerning automated decision-making is not disputed in this paper. Rather our concern has been to explicate the fundamental mismatch between legal and ML notions of explanation. While it is clear that what constitutes explanation in either domain is contested, the legal requirement that automated decision-making be justified **does not equate** to providing a causal account as to how an algorithmic model arrived at a decision. There is **no** general provision in GDPR for *ex post* explanations of specific decisions arrived at by algorithmic machines as these are only warranted in certain circumstances, and then what is mandated is that a human being make the decision arrived at available to normative assessment. Being accountable in law does not require causal explanation, though it may draw on causal accounts if they can be provided.

At the heart of the debate concerning the right to an explanation is a fundamental misreading that has consequences for the machine learning community and society at large. The misreading creates the unrealistic and unrealisable expectation that what is required from the ML community is machines whose inner workings are intelligible to society, and amongst society that this is indeed what will be delivered. However, not only is the need for explanation *only* mandated in situations where decision-making is solely or wholly automated, the kind of explanation offered by ML methods of interpretability can only play a limited role in enabling the legally defensible explanations required by GDPR, assuming that the problems of inscrutability and non-intuitiveness can be dealt with and that is by no means given. Indeed the increasing complexity of algorithmic machines, which underpins their value, would seem to mitigate against that possibility to a significant extent in the foreseeable future. Even then, causal accounts will not be sufficient and external explanations that account for why decisions happen to be made in the ways that are, and whether or not they are reasonable, will have to provided and satisfied.

Preoccupation with the right to an explanation creates the misleading impression that ML, in collaboration with ethics and law, will be able to address the social imperative of accommodating AI in everyday life. However, society at large will not accept AI solely on the basis of ethical assertions or legal assurances, everyday life does not work that way. We must face up to the fact that explanations come to an end and we have at some point and to some greater or (in due course perhaps) lesser extent to live with inscrutable and non-intuitive machines. Understanding how we might accommodate technology that defies intuition and resists comprehension, and the challenges that this in turn creates for machine learning and AI, is a critical imperative. In introducing this paper we drew on Lipton's (2016) observation that, *"as machine learning continues to exert influence upon society, we must be sure that we are solving the right problems."* We recommend that the right to an explanation be considered harmful then, as it misdirects us and diverts our attention away from solving problem of accommodation. Explanation does not hold the key to this.

This is not to dismiss efforts within machine learning to make algorithmic machines more intelligible; clearly we have need as human beings to understand as much as we can of the machines that we build and their operations (but not all of us do and, as GDPR makes apparent, only some of us do some of the time). Nor is it to say that ML should pay no heed to the requirements of law; GDPR makes it clear that automated decision-making will be held accountable and ML methods of interpretability may support data controllers in this. It is, however, to recognise the limits of explanation: of what *can* be accounted for by ML methods of interpretability in the face of inscrutability and non-intuitiveness and what *more* needs to be taken into account than explanation provides for AI to find its way into our everyday lives. The social imperative of accommodation requires that we reach beyond the ELT matrix to involve new forms of societal representation and new narratives that enable us to shape new kinds of "interface" to the algorithmic machine; interfaces that go beyond explanation and allow us to situate AI in context and make it at home in our everyday lives.

Acknowledgements

This work was supported by the Engineering and Physical Sciences Research Council [grants EP/M001636/1, EP/M02315X/1].

References

A29WP WP248. *Guidelines on Data Protection Impact Assessment (DPIA)*. European Commission, 13 October, 2017.

A29WP WP251. *Guidelines on Automated Individual Decision-making and Profiling for the Purposes of Regulation 2016/679*. European Commission, 22 August, 2018.

Budish, R., Bavitz, C., Doshi-Velez, F., Gershman, S., Kortz, M., O'Brien, D., Shieber, S., Waldo, J., Weinberger, D., and Wood, A. *Accountability of AI Under the Law: The Role of Explanation*. arXiv:1711.01134, 21 November, 2017.

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., and Elhadad, N. Intelligible models for healthcare: predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1721–1730, Sydney, NSW, 2015. ACM Press.

Chen, J. The Dangers of Accuracy: Exploring the Other Side of the Data Quality Principle. *European Data Protection Law Review*, 4 (1): 36–52, 2018.

Diakopoulos, N. Accountability in algorithmic decision-making. *Communications of the ACM*, 59 (2): 56-62, 2016.

Edwards, L., and Veale, M. Slave to the Algorithm? Why a 'right to an explanation' is probably not the remedy you are looking for. *Duke Law & Technology Review*, 16 (1): 18–84, 2017.

Floridi, L, Cave, J., Davis, J., Mittelstadt, B., Raab, C., Wachter, S., Weller A., and Maskell, R. (2017) *Written evidence submitted by The Alan Turing Institute (ALG0073)*. House of Commons' Science and Technology Committee, 28 February, 2017.

GDPR. Regulation 2016/679 General Data Protection Regulation. *Official Journal of the European Union*, 59: 1–149, 2016.

Goodman, B., and Flaxman, S. EU regulations on algorithmic decision-making and "a right to an explanation". In Kim, B., Malioutov, M. and Varshney, K.R. (eds.), *Proceedings of the ICML Workshop on Human Interpretability in Machine Learning (WHI 2016)*, pp. 26–30, New York City, NY, 2016. International Machine Learning Society.

Guidotti, R, Monreale, A., Ruggieri, S., Turini F., Pedreschi, D., and Giannotti, F. A survey of methods for explaining black box models. *ACM Computing Surveys*, 51 (5): Article No. 93, 2018.

Gunning, D. *Explainable Artificial Intelligence (XAI)*. DARPA, 2018.

ICO. *GDPR Articles and Recitals*, 2018. https://ico.org.uk/media/about-the-ico/disclosure-log/2014536/irq0680151-disclosure.pdf

Kuijer, L., and Giaccardi, E. (2018) Co-performance: conceptualising the role of artificial agency in the design of everyday life. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, paper no. 125, Montreal, Canada, 2018. ACM Press.

Knight, W. (2017) The dark secret at the heart of AI. *MIT Technology Review*, 11 April, 2017.

Levin, S., and Carrie, J. Self-driving Uber kills Arizona woman in first fatal crash involving pedestrian. *The Guardian*, 19 March, 2018.

Lie, M & Sørensen, K.H. (eds.). *Making Technology Our Own? Domesticating Technology into Everyday Life*. Scandinavian University Press, Oslo, 1996.

Lindley, J., Coulton, P., and Sturdee, M. Implications for adoption. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 265–277, Denver, CO, 2017. ACM Press.

Lipton, Z.C. (2016) The mythos of model interpretability. In Kim, B., Malioutov, M. and Varshney, K.R. (eds.), *Proceedings of the ICML Workshop on Human Interpretability in Machine Learning (WHI 2016)*, pp. 96–100, 2016. International Machine Learning Society.

Malcolm, N. *Nothing is Hidden: Wittgenstein's Criticism of his Early Thought*. Basil Blackwell, Oxford, UK, 1986.

Medsker, L. Algorithms and algorithmic transparency. *AI Matters: A Newsletter of ACM SIGAI*, 2 August, 2017.

Mittelstadt, B. Auditing for transparency in content personalization systems. *International Journal of Communication*, 10: pp. 4991–5002, 2016.

Molnar, C. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*, 2018. https://christophm.github.io/interpretable-ml-book.

Nilsson, T., Crabtree, A., Fischer, J.E., Koleva, B. Breaching the future: understanding human challenges of autonomous systems for the home. *Social Science Research Network*, 10 August, 2018.

Pasquale, F. *The Black Box Society: The Secret Algorithms that Control Money and Information*. Harvard University Press, Cambridge, MA, 2015.

Porcheron, M., Fischer, J.E., Reeves, S., and Sharples, S. Voice interfaces in everyday life. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, paper no. 640, Montreal, Canada, 2018. ACM Press.

Raz, J. Reasons: Explanatory and Normative. *From Normativity to Responsibility* (ed. Raz, J.), pp. 13–35. Oxford University Press, Oxford, UK, 2011.

Ribeiro, M.T., Singh, S., and Guestrin, C. "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144, San Francisco, CA, 2016.

Royal Society. *Machine Learning: The Power and Promise of Computers that Learn by Example*, April 2017.

Selbst, A.D., and Barocas, S. The intuitive appeal of explainable machines. *Fordham Law Review*, 87 (3): 1085–1139, 2018.

Selbst, A.D., and Powles, J. Meaningful information and the right to an explanation. *International Data Privacy Law*, 7 (4): 233–242, 2017.

Shegloff, E.A. Introduction by Emanuel A. Schegloff. *Harvey Sacks Lectures on Conversation, Volume I* (ed. Jefferson, G.), pp. ix–lxii. Blackwell Publishing, Malden, MA, 1992.

Queloz, M. Two orders of things: Wittgenstein on reasons and causes. *Philosophy*, 92 (3): 369–397, 2017.

Urquhart, L., Lodge, T., and Crabtree, A. *Demonstrably doing accountability in the Internet of Things*. arXiv:1801.07168, 22 January, 2018.

Wachter, S., Mittelstadt, B., and Floridi, L. Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation. *International Data Privacy Law*, 7 (2): 76–99, 2017.

Walport, M. The rise of machines: are algorithms sprawling out of our control? *Wired*, 1 April, 2017.

Winch, P. (1958) *The Idea of a Social Science*. Routledge & Kegan Paul, Great Britain, 1958.

Wittgenstein, L. *Philosophical Investigations*. Blackwell Publishers, Oxford, UK, 1992.

Zoldi, S. How and why AI will evolve in 2018. *Compare the Cloud*, 9 January, 2018.